On-line clustering of individual sequences (or Prediction with several strategies)

Sébastien Loustau*

February 7, 2014

Abstract

We know that ℓ_0-penalized methods have good theoretical properties but unfortunately high computational cost. On the contrary, convex relaxations - such as the Lasso - have been introduced but their theoretical guarantees hold for restricted models. To tackle this impasse, [13, 14] come up with sparsity priors in a Pac-Bayesian framework. They give rise to good theoretical properties, i.e. sparsity oracle inequalities, reached by computationally attractive sequential procedures. In this paper, we investigate this issue in clustering.

We construct online *clustering* algorithms which learn according to the following game protocol. At each trial $t \geq 1$, nature reveals a deterministic $x_t \in \mathbb{R}^d$, $d \geq 1$. A forecaster predicts the next value with several - and as small as possible - proposals. Then, nature reveals the next value and the forecaster pays the minimal distance between this value and its set of proposals. To deal with this problem, we use the Pac-Bayesian theory with group-sparsity priors. It gives sparsity regret bounds and allows us to perform online clustering of a possible non-stationnary process, without any knowledge about the number of clusters. These results can be applied to the classical i.i.d. case to deal with the problem of model selection clustering as well as high dimensional clustering.

1 Introduction

Prediction of individual sequences is the core of a huge amount of work these last decades in game theory and statistics. The problem could be summarized as follows. A blackbox reveals at each trial t a real value $x_t \in \mathbb{R}$, which could be a temperature at a given time, a risk asset, or an unemployment rate. Then, a forecaster predicts the next value based on the past observations and expert advices. These expert advices could be based on deterministic - or stochastic - models, or even adversarial. The goal is to predict as well as the best expert, no matter what sequence is produced by the blackbox. This sequential game has been investigated by many authors. We can mention the monograph of [9] for a nice introduction to the area (see also the pioneering work of [22]). Very often, the introduced algorithms are based on convex combinations of expert advices, where coefficients depend on the past performances of each expert. In this paper, we suggest to tackle a more difficult task by considering *vector*-valued instances $x_t \in \mathbb{R}^d$, $d \geq 1$, and *no* expert advice.

Instead, we construct online *clustering* algorithms which learn according to the following protocol. On each day t, the forecaster must predict the next instance $x_t \in \mathbb{R}^d$ with at most $p \geq 1$ possible "proposals" or "strategies". On the morning of day t, he has access to the inputs $x_1, \dots, x_{t-1}$ of the previous days. Based on these instances, he must propose a codebook of $p \geq 1$ strategies $\hat{\mathbf{c}}_t = (\hat{c}_{t,1}, \dots, \hat{c}_{t,p}) \in \mathbb{R}^{dp}$. At the end of the day, he receives x_t and encurs a loss - or distortion - $\ell(\hat{\mathbf{c}}_t, x_t)$, where:

$$\ell(\hat{\mathbf{c}}_t, x_t) = \min_{j=1,\dots,p} |\hat{c}_{t,j} - x_t|_2^2, \tag{1.1}$$

*Université d'Angers, LAREMA, loustau@math.univ-angers.fr

and $|\cdot|_2$ stands for the Euclidean norm in $\mathbb{R}^d$. The goal of the forecaster is to control the cumulative distortion $\sum_{t=1}^T \ell(\hat{\mathbf{c}}_t, x_t)$, with $|\hat{\mathbf{c}}_t|_0$ as small as possible, where $|\hat{\mathbf{c}}_t|_0$ corresponds to the number of non-zero strategies at time t, i.e.:

$$|\mathbf{c}|_0 := \text{card}\{j = 1, \ldots, p : c_j \neq (0, \ldots, 0)^\top \in \mathbb{R}^d\}, \ \forall \mathbf{c} = (c_1, \ldots, c_p) \in \mathbb{R}^{dp}.$$

At this stage, it is important to explain what means $|\hat{\mathbf{c}}_t|_0$ as small as possible. Firstly, note that a possible candidate strategy - but out of interest in practice - is the following. At each trial $t \geq 1$, the forecaster puts a proposal on each past instance $x_t \in \mathbb{R}^d$ and let the other components to zero. This trivial system will have the property $|\hat{\mathbf{c}}_t|_0 = t$, for any $t \geq 1$. Consequently, this strategy will induce small cumulative loss $\sum_{t=1}^T \ell(\hat{\mathbf{c}}_t, x_t)$ but huge complexity, and is equivalent to the so-called "overfitting phenomenon". In this contribution, we want to develop algorithms which summarize the information of the deterministic sequence, namely such that $|\hat{\mathbf{c}}_t|_0 << t$.

A motivating example is as follows. A t-shirt sailer receives online data about their sales, customer after customer (such as prize, color and shape). After each checkout process t, he must predict the next instance in order to market appropriately to the current customer's patterns. Since different social clusters are involved (such as boys, teens, or gothics), he can advise different strategies or attempts. The only restriction he has is to come up with a finite - and as small as possible - number of strategies in order to summarize the demand (he has not access to an infinite store size). Moreover, since fashion changes over time, the retailer wants to learn in an online way.

In this contribution, we make no assumption about the sequence of inputs that arrives at each trial. Our results hold for a worst case sequence of instances. It allows to tackle non stationarity in the learning process and predict - or cluster - sequence of trials with time-varying structure. This problem has been, as far as we know, very poorly treated in the literature. In the framework of expert advice, we can mention [11], where different clustering algorithms are aggregated at each trial to get an online clustering algorithm. [31] investigates an online version of the spherical k-means algorithm. In the present paper we have not any expert advice (such as k-means). In particular, we have no idea about the number of clusters to use in the learning protocol.

From theoretical viewpoint, we are interested in sparsity regret bounds introduced in [16]. More precisely, we recommend to control the cumulative loss as follows:

$$\sum_{t=1}^T \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{\mathbf{c} \in \mathbb{R}^{dM}} \left\{ \sum_{t=1}^T \ell(\mathbf{c}, x_t) + \lambda |\mathbf{c}|_0 \right\} + r_\lambda(T), \tag{1.2}$$

where $r_\lambda(T)$ is a residual term and $\lambda > 0$ is a temperature parameter. It has to be calibrated in order to minimize the right hand side of (1.2). In other words, we want to control the regret of our sequential procedure to have not reach the compromise between fitting the data and compress the information (i.e. the infimum which appears in the right hand side). Going back to the t-shirt retailer example, it means that we are looking at a strategy that fits to the customer's patterns as well as possible, but also which minimizes the number of offers. This compromise is of first interest in information theory and statistics.

To get sparsity regret bounds (1.2), we use Pac-Bayesian bounds in the spirit of [8]. We are also largely inspired from the seminal paper of [2], which presents a unified Pac-Bayesian theory for both the deterministic (or worst case) scenario and the batch (or i.i.d.) case. The second ingredient of our results in the introduction of new sparsity priors. It is based on previous priors recommended in prediction under the sparsity assumption in high dimensional statistics (see for instance [12, 13]). The main challenge in this area is to introduce procedures which balance good theoretical properties (such as ℓ_0-penalized estimators) and low computational cost (such as solutions of convex or linear programming). In the framework of individual sequences, this issue has been considered recently in [16] in the problem of online linear regression. From a theoretical perspective, the present contribution develops sparsity regret bounds in online clustering.

Our algorithms are based on standard sequential randomized procedures largely inspired from the literature cited above. It has been noted in [14] that these methods are computationally feasible for relatively large dimensions of the problems, by using a so-called Langevin Monte-Carlo method. These computational aspects have been also considered in [1] in a sparse single index model and in [18] in a sparse additive model. Note that this direction is of first interest in clustering since the available algorithms are not quiet satisfactory. In the standard i.i.d. case, it has been noticed by many authors that the existing methods, such as the popular k-means, suffers from non-convexity. As a result, the initialization affects the performances of the algorithm, which does not guarantee the convergence to the global minimum of the empirical risk (see [7]). Within the PAC-Bayesian framework of this paper, up to some Monte-Carlo approximation, we are abble to compile the algorithm which have the desire theoretical properties. Moreover, in this work, we enlarge the framework to non-stationary processes in comparison to the i.i.d. case. That's why the algorithmic part is an interesting direction for future works.

The paper is organized as follows. In Section 2, we present the sequential randomized algorithm and give the main results of the paper, for the problem of online clustering with known horizon T. The online clustering algorithm reaches sparsity regret bounds as in (1.2). Then, in Section 3, we turn out into the problem of adaptation, namely the knowledge of T. We give an adaptive version of the online algorithm of Section 2, where the temperature parameter $\lambda > 0$ could vary over time. Finally, we illustrate in Section 4 the power of the Pac-Bayesian theory in the standard i.i.d. case to investigate model selection clustering as well as high dimensional clustering. We conclude in Section 5 with a short discussion whereas Section 6 is dedicated to the proofs of the main results.

2 Main results

In this section, we describe the first online clustering algorithm and its generalization ability. We begin with a general Pac-Bayesian bound, which holds for any prior and any temperature parameter. This result allows us to get a sparsity regret bound for our problem, by using a group-sparsity prior.

2.1 The algorithm

For any integer $d, p \geq 1$, we denote by $\mathcal{M}_1^+(\mathbb{R}^{dp})$ the set of probability measure on $\mathbb{R}^{dp}$. We start by describing the algorithm as follows. Let us introduce a prior $\pi \in \mathcal{M}_1^+(\mathbb{R}^{dp})$ and an inverse temperature parameter $\lambda > 0$. At the beginning of the game, we draw $\hat{\mathbf{c}}_1$ with law $\hat{p}_1 := \pi$. We fix $S_0 \equiv 0$. Then, learning proceeds as the following sequence of trials $t = 1, \ldots, T$:

- Get x_t and compute $S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c}, x_t) + \frac{\lambda}{2}[\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2, \ \forall \mathbf{c} \in \mathbb{R}^{dp}$.
- Let $\hat{p}_{t+1}(d\mathbf{c}) := \frac{e^{-\lambda S_t(\mathbf{c})}}{W_t}\pi(d\mathbf{c}) \in \mathcal{M}_1^+(\mathbb{R}^{dp})$, where $W_t = \mathbb{E}_{\mathbf{c}\sim\pi} e^{-\lambda S_t(\mathbf{c})}$.
- Draw $\hat{\mathbf{c}}_{t+1}$ according to the law $\hat{p}_{t+1}$.

Then, we have constructed a vector of probability measures $(\hat{p}_1, \ldots, \hat{p}_{T+1})$, where each $\hat{p}_t \in \mathcal{M}_1^+(\mathbb{R}^{dp})$ is calculated thanks to the sequence of past instances $x_1, \ldots, x_{t-1}$ and the realizations of $(\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_{t-1})$. More precisely, the principle is to update the current error of any codebook $\mathbf{c} \in \mathbb{R}^{dp}$ as follows:

$$S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c}, x_t) + \frac{\lambda}{2}[\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2, \tag{2.1}$$

where $\lambda > 0$ is some temperature parameter. At each trial t, the loss of a codebook is decomposed as the loss over the past, the current loss $\ell(\mathbf{c}, x_t)$ and a stability term that ensures $\hat{\mathbf{c}}_{t+1}$ to be not

so far from $\hat{\mathbf{c}}_t$. This term can be viewed as a penalization term to better control the variance in our procedure (see [2] for details and inequality (2.4) below). Due to the construction of a randomized estimator $\hat{\mathbf{c}}$, we are interested in the cumulative expected loss, given by:

$$\mathcal{E}_T(\hat{\mathbf{c}}) := \sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1,\ldots,\hat{p}_t)} \ell(\hat{\mathbf{c}}_t, x_t), \tag{2.2}$$

where for each $t \geq 1$, the product measure $(\hat{p}_1, \ldots, \hat{p}_t)$ is constructed in the algorithm.

2.2 A Pac-Bayesian bound

Pac-Bayesian bounds go back to the work of [23] (see also [8] or more recently [27]). It gives a control in expectation of the risk of any randomized estimator. The precise expression of the upper bounds depends on the context, but it is very often an empirical risk penalized in terms of Kullback-Leibler divergence. In what follows, the Kullback-Leibler divergence of two measures $\rho, \pi \in \mathcal{M}_1^+(\mathbb{R}^{dp})$ in the measurable space $(\mathbb{R}^{dp}, \mathcal{B}(\mathbb{R}^{dp}))$ is defined as:

$$\mathcal{K}(\rho,\pi) := \begin{cases} \mathbb{E}_{\mathbf{c}\sim\rho} \log \frac{d\rho}{d\pi}(g) & if \rho << \pi, \\ +\infty & \text{otherwise.} \end{cases}$$

A nice property (see for instance [8]) is the following duality formula. For any measurable function $h : \mathbb{R}^{dp} \to \mathbb{R}$, we have:

$$\log \mathbb{E}_{\mathbf{c}\sim\pi} e^{h(\mathbf{c})} = \sup_{\rho \in \mathcal{M}_1^+(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\mathbf{c}\sim\rho} h(\mathbf{c}) - \mathcal{K}(\rho,\pi) \right\}. \tag{2.3}$$

Since the earlier work of Mac Allester, many authors have investigated Pac-Bayesian bounds. For our purpose, we can mention [2], which has largely inspired the result of Theorem 1 below. In particular, the construction of the algorithm in Section 2.1 - and more precisely the update of the current error described in (2.1) - ensures the following property:

$$\forall \lambda > 0, \forall \rho \in \mathcal{M}_1^+(\mathbb{R}^{dp}), \forall x \in \mathbb{R}^d,\ \mathbb{E}_{\mathbf{c}'\sim\rho} \ell(\mathbf{c}', x) \leq -\frac{1}{\lambda} \mathbb{E}_{\mathbf{c}'\sim\rho} \log \mathbb{E}_{\mathbf{c}\sim\rho} e^{-\lambda(\ell(\mathbf{c},x) + \frac{\lambda}{2}[\ell(\mathbf{c},x_t) - \ell(\mathbf{c}',x_t)]^2)}. \tag{2.4}$$

The assertion (2.4) is proved in [2] in a quiet general setting and called the variance inequality. This inequality can be traced back to [19] (see also [20] in the i.i.d. setting). In our framework, it is the starting point to get the following result:

Theorem 1 *For any deterministic sequence $(x_t)_{t=1}^T \in \mathbb{R}^{dT}$, for any $p \in \mathbb{N}^\star$, any $\lambda > 0$ and any prior $\pi \in \mathcal{M}_+(\mathbb{R}^{dp})$, the cummulative loss (2.2) satisfies:*

$$\mathcal{E}_T(\hat{\mathbf{c}}) \leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\boldsymbol{c}\sim\rho} \sum_{t=1}^{T} \ell(\boldsymbol{c}, x_t) + \frac{\mathcal{K}(\rho,\pi)}{\lambda} + \frac{\lambda}{2} \mathbb{E}_{(\hat{p}_1,\ldots,\hat{p}_T)} \mathbb{E}_{\boldsymbol{c}\sim\rho} \sum_{t=1}^{T} [\ell(\boldsymbol{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2 \right\}. \tag{2.5}$$

The bound of Theorem 1 gives a control of the expected cumulative loss of the randomized procedure described in Section 2.1. The interesting point with Theorem 1 is that it holds for any choice of prior π, as well as any inverse temperature parameter $\lambda > 0$. It allows us in the sequel to choose a suitable prior, namely a group-sparsity prior, to give a sparsity regret bound for our problem.

Proof of Theorem 1. From (2.4), we have, for any $\rho \in \mathcal{M}_1^+(\mathbb{R}^{dp})$, for any $\lambda > 0$:

$$\forall x \in \mathbb{R}^d,\ \mathbb{E}_{\mathbf{c}'\sim\rho} \ell(\mathbf{c}', x) \leq -\frac{1}{\lambda} \mathbb{E}_{\mathbf{c}'\sim\rho} \log \mathbb{E}_{\mathbf{c}\sim\rho} e^{-\lambda[\ell(\mathbf{c},x) + \frac{\lambda}{2}(\ell(\mathbf{c},x) - \ell(\mathbf{c}',x))^2]}.$$

Then, for any $t \in \{1, \dots T\}$, for fixed $(\hat{\mathbf{c}}_1, \dots, \hat{\mathbf{c}}_{t-1})$, we have:

$$\mathbb{E}_{\mathbf{c}' \sim \hat{p}_t} \ell(\mathbf{c}', x_t) \leq -\frac{1}{\lambda} \mathbb{E}_{\mathbf{c}' \sim \hat{p}_t} \log \mathbb{E}_{\mathbf{c} \sim \hat{p}_t} e^{-\lambda[\ell(\mathbf{c}, x_t) + \frac{\lambda}{2}(\ell(\mathbf{c}, x) - \ell(\mathbf{c}', x))^2]}.$$

Integrating with respect to $(\hat{p}_1, \dots, \hat{p}_{t-1})$ and summing over t, we get:

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq -\frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_t)} \log \mathbb{E}_{\mathbf{c} \sim \hat{p}_t} e^{-\lambda[\ell(\mathbf{c}, x_t) + \frac{\lambda}{2}(\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t))^2]} =: \Delta_T.$$

Next step is to rewrite the RHS Δ_T. By construction of the algorithm and namely the equality $S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c}, x_t) + \frac{\lambda}{2}(\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t))^2$, we have:

$$\begin{aligned}
\Delta_T &= -\frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_t)} \log \mathbb{E}_{\mathbf{c} \sim \hat{p}_t} e^{-\lambda[S_t(\mathbf{c}) - S_{t-1}(\mathbf{c})]} \\
&= -\frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_t)} \log \left(\frac{\int_{\mathbb{R}^{dp}} e^{-\lambda[S_t(\mathbf{c}) - S_{t-1}(\mathbf{c})]} e^{-\lambda S_{t-1}(\mathbf{c})} d\pi(\mathbf{c})}{W_{t-1}} \right) \\
&= -\frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_t)} \log \left(\frac{W_t}{W_{t-1}} \right) = -\frac{1}{\lambda} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_T)} \log \prod_{t=1}^{T} \left(\frac{W_t}{W_{t-1}} \right).
\end{aligned}$$

We are now on time to apply the chain rule (see [4]) to get:

$$\begin{aligned}
\Delta_T &= -\frac{1}{\lambda} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_T)} \log (W_T) \\
&= -\frac{1}{\lambda} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_T)} \log \mathbb{E}_{\mathbf{c} \sim \pi} e^{-\lambda S_T(\mathbf{c})} \\
&= -\frac{1}{\lambda} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_T)} \sup_{\rho \in \mathcal{M}_1^+(\mathbb{R}^{dp})} \{-\lambda \mathbb{E}_{\mathbf{c} \sim \rho} S_T(\mathbf{c}) - \mathcal{K}(\rho, \pi)\} \\
&= \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_T)} \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\mathbf{c} \sim \rho} S_T(\mathbf{c}) + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\} \\
&\leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\mathbf{c} \sim \rho} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \frac{\mathcal{K}(\rho, \pi)}{\lambda} + \frac{\lambda}{2} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_T)} \mathbb{E}_{\mathbf{c} \sim \rho} \sum_{t=1}^{T} [\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2 \right\},
\end{aligned}$$

where we use the Kullback duality formula (2.3) at the third line.

□

2.3 A group-sparsity prior

Group-sparsity encourages occurences of whole blocks of zeros in the decision vector (see [29]). It has been used in many applications, such as genetics or image annotation (see [30]), where the Lasso is not consistent for variable selection in high correlation settings. In this paper, we are looking at a vector $\mathbf{c} = (c_1, \dots, c_p) \in \mathbb{R}^{dp}$ such that $|\mathbf{c}|_0 := \text{card}\{j = 1, \dots, p : c_j \neq (0, \dots, 0)^\top\}$ is small, namely a so-called group-sparsity. More precisely, we want that many c_j in $\mathbf{c} = (c_1, \dots c_p)$ are $(0, \dots, 0)^\top$ of $\mathbb{R}^d$. To deal with this issue, we come up with a new kind of prior, called a group-sparsity prior. It consists of a product of multivariate Student's distribution $\sqrt{2}\tau T_d(3)$, where $\tau > 0$ is a scaling parameter and $T_d(3)$ is the d-multivariate Student with three degrees of freedom. It can be viewed as a generalization of the prior used in [13], where a product of univariate Student is considered. Essentially, a group-sparsity prior generalizes a sparsity prior with groups of size $d \geq 1$, instead of groups of size 1. Consequently, we use the multivariate Student's distribution presented in [21], defined as the ratio between a gaussian vector and the

square root of an independent χ^2 distribution with 3 degrees of freedom. In our case, it leads to the following representation:

$$\pi_S(d\mathbf{c}) := \prod_{j=1}^{p} \left\{ C_{R,\tau}^{-1} \left(1 + \frac{|c_j|_2^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \mathbb{I}(|c_j|_2 \leq 2R) \right\} d\mathbf{c}, \tag{2.6}$$

where $C_{R,\tau} := c_{d,\tau}\mathbb{P}(\sqrt{2}\tau T_d(3) \in \mathcal{B}_2(2R))$ for some constant $c_{d,\tau} > 0$. Here, $R > 0$ is a threshold that could be chosen arbitrarily big. Roughly speaking, the scaling parameter $\tau > 0$ - which can be fixed to a really small parameter - ensures sparsity for the vector of p groups $\sqrt{2}\tau T_d(3)$ whereas the heavy tails property of $T_d(3)$ guarantees that a small proportion of groups are quiet far from zero. From theoretical viewpoint, the introduction of the group-sparsity prior (2.6) gives rise to the following lemma:

Lemma 1 *Let $p \in \mathbb{N}^\star$, $\tau, R > 0$ and π_S defined in (2.6). Let $\mathbf{c} = (c_1, \ldots, c_p) \in \mathbb{R}^{dp}$ such that $c_j \in \mathcal{B}_2(R) = \{c \in \mathbb{R}^d : |c|_2 \leq R\}$, for any $j \in \{1, \ldots, p\}$. Introduce p_0 the following translated version of π_S with mean $\mathbf{c}$:*

$$p_0(d\mathbf{c}') = \prod_{j=1}^{p} \left\{ C_{R,\tau}^{'-1} \left(1 + \frac{|c_j' - c_j|_2^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \mathbb{I}(|c_j' - c_j|_2 \leq R) \right\} d\mathbf{c}',$$

where here $C'_{R,\tau} := c_{d,\tau}\mathbb{P}(\sqrt{2}\tau T_d(3) \in \mathcal{B}_2(R))$. Then $p_0 << \pi_S$ and we have:

$$\begin{aligned}
\mathcal{K}(p_0, \pi_S) &\leq 2\sum_{j=1}^{p} \log\left(1 + \frac{|c_j|_2}{\sqrt{6}\tau}\right) + p\log\left(\frac{C_{R,\tau}}{C'_{R,\tau}}\right) \\
&\leq 2|\mathbf{c}|_0 \log\left(1 + \frac{\sum_{j=1}^{p}|c_j|_2}{\sqrt{6}\tau|\mathbf{c}|_0}\right) + \frac{12pd\tau^2}{R^2}.
\end{aligned}$$

The proof of the lemma is postponed to Section 6. It is based on the result of [13] and the property of the multivariate Student's distribution (see [21]). In particular, it is important to stress that the multivariate Student distribution $T_d(3)$ introduced in [21] is not a product of independent univariate Student's distribution.

Note that the first inequality of Lemma 1 shows that we can also consider approximate group-sparsity, i.e. codebook $\mathbf{c} = (c_1, \ldots, c_p)$ where many c_j, $j = 1, \ldots, p$ are very close to zero. However, we state the main results with the second inequality of Lemma 1, which leads to (exact) sparsity regret bounds in our setting.

2.4 Sparsity regret bounds

In this paragraph, we state the main results of this section, i.e. sparsity regret bounds of the form (1.2) for the algorithm described in Section 2.1. The first result is a direct consequence of Theorem 1 and the introduction of the sparsity prior (2.6).

Corollary 1 *For any deterministic sequence $(x_t)_{t=1}^T$, any $\tau, \lambda, R > 0$, let us consider the algorithm of Section 2.1 using prior π_S defined in (2.6). Then, the following holds:*

$$\begin{aligned}
\mathcal{E}_T(\hat{\mathbf{c}}) \leq \inf_{\mathbf{c}\in\mathbb{R}^{dp}:\forall j, |c_j|_2 \leq R} &\left\{ \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \frac{2|\mathbf{c}|_0}{\lambda} \log\left(1 + \frac{\sum_{j=1}^{p}|c_j|_2}{\sqrt{6}|\mathbf{c}|_0\tau}\right) \right\} \\
&+ CT\left(\sqrt{6p}\tau + 2C\lambda R^2\right) + \frac{12pd\tau^2}{\lambda R^2},
\end{aligned}$$

where $C := 2B_T + 3R$ with $B_T = \max_{t=1,\ldots,T}|x_t|_2$.

The proof is a direct consequence of Theorem 1, gathering with Lemma 1 above. It is postponed to Section 6 for concision. Moreover, by tuning the couple of parameters (τ, λ) in the algorithm, we can show the following corollary.

Corollary 2 *For any deterministic sequence $(x_t)_{t=1}^T$, any $R > 0$, let us consider the algorithm of Section 2.1 using prior π_S defined in (2.6) and parameters $(\tau, \lambda) = \left((pT)^{-1/2}, T^{-1/2}\right)$. Then, the following holds:*

$$\mathcal{E}_T(\hat{\boldsymbol{c}}) \leq \inf_{\boldsymbol{c}\in\mathbb{R}^{dp}:\forall j, |c_j|_2 \leq R} \left\{ \sum_{t=1}^T \ell(\boldsymbol{c}, x_t) + 2|\boldsymbol{c}|_0 \sqrt{T} \log\left(1 + \frac{\sqrt{pT}\sum_{j=1}^p |c_j|_2}{\sqrt{6}|\boldsymbol{c}|_0}\right)\right\} + \sqrt{T}\left(\sqrt{6}C + 2C^2R^2\right) + \frac{12d}{\sqrt{T}R^2}.$$

The choice of (τ, λ) in Corollary 2 gives rise to a sparsity regret bound with rate $\mathcal{O}(\sqrt{T})$, up to a $\log T$ factor. Fix $p = T$ and suppose the infimum of the RHS is reached by a codebook $\mathbf{c}^\star$ such that $|\mathbf{c}^\star|_0 = s$ for some sparsity index $s \in \mathbb{N}^\star$. Then, we have the following bound:

$$\mathcal{E}_T(\hat{\mathbf{c}}) - \mathcal{E}_T(\mathbf{c}^\star) \lesssim s\sqrt{T}\log T,$$

where $a \lesssim b$ means that there exists a constant $c > 0$ such that $a \leq cb$.
Note that the residual term of Corollary 2 when $\tau := \tau(p, T)$ does not depend on p. As a result, we can choose $p = T$ in the algorithm without any influence on the rate of convergence. However, the choice of the couple (λ, τ) depends explicitly on the horizon T, which is not known in a pure online setting. This problem is considered in Section 3 where an adaptive version of the algorithm of Section 2.1 is given.
Finally, the infimum in Corollary 1 and Corollary 2 is restricted to $\{\mathbf{c} \in \mathbb{R}^{dp} : \forall j,\ |c_j|_2 \leq R\}$. This arises for technicalities in the proof and could be extend to the whole space $\mathbb{R}^{dp}$. In the spirit of [16], we can consider the truncated loss function:

$$\ell(\mathbf{c}, x) = \min_{j=1,\dots,p} |x - [c_j]_B|_2^2,$$

where $[c]_B$ is the projection of c into the ball of radius $B > 0$ in $\mathbb{R}^d$ defined as:

$$[c]_B = c\ \mathbb{1}(|c|_2 \leq B) + \arg\min_{u\in\mathbb{R}^d:|u|_2\leq B} |c - u|_2^2.$$

This step is not of major importance since by choosing large enough constant $B > 0$, it has no influence on the loss function.

3 Adaptation

The choice of the inverse temperature $\lambda > 0$ in Corollary 2 depends explicitly on the horizon T of the deterministic sequence. However, if we consider a pure online setting, the size of the deterministic sequence is unknown. This problem is called adaptation in the deterministic literature and has been extensively studied in the context of prediction with expert advices (see [3, 10, 16]). Originally, one can use a doubling trick, which consists in restarting the algorithm at periods of exponentially increasing lengths of size 2^k, for $k \geq 1$. A more natural alternative is to let the tuning parameters depend on the trial $t \geq 1$. The idea has been introduced in [3] and influences the regret bounds by only a constant factor. This approach is developed below.

The adaptive online clustering algorithm mimics the previous sequential procedure as follows. Let us consider a prior $\pi \in \mathcal{M}_1^+(\mathbb{R}^{dp})$ and a non-increasing sequence of temperature parameter $(\lambda_t)_{t=1}^{T+1}$. At the beginning of the game, we draw $\hat{\mathbf{c}}_1$ with law $\hat{p}_1 := \pi$. We fix $S_0 \equiv 0$. Then, learning proceeds as the following sequence of trials $t \in \{1, \dots, T\}$:

- Get x_t and compute: $S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c}, x_t) + \frac{\lambda_t}{2}[\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2, \ \forall \mathbf{c} \in \mathbb{R}^{dp}$.
- Let $\hat{p}_{t+1}(d\mathbf{c}) := \frac{e^{-\lambda_{t+1} S_t(\mathbf{c})}}{W_t} \pi(d\mathbf{c})$ where $W_t = \mathbb{E}_{\mathbf{c}\sim\pi} e^{-\lambda_{t+1} S_t(\mathbf{c})}$.
- Draw $\hat{\mathbf{c}}_{t+1}$ according to the law $\hat{p}_{t+1}$.

The adaptive algorithm presented above lead to a randomized estimator denoted as $\hat{\mathbf{c}}_{\text{adapt}}$. It depends on a sequence of non-increasing temperature parameters $(\lambda_t)_{t=1}^{T+1}$. From Section 2, the choice of $\lambda_t = 1/\sqrt{t}$ seems optimal and gives rise to the following adaptive regret bound.

Theorem 2 *For any deterministic sequence $(x_t)_{t=1}^T$, any $\tau, R > 0$, let us consider the adaptive algorithm with $\lambda_t = 1/\sqrt{t}$, for $t = 1, \ldots, T+1$ and prior π_S defined in (2.6). Then:*

$$\mathcal{E}_T(\hat{\mathbf{c}}_{\text{adapt}}) \leq \inf_{\boldsymbol{c}\in\mathbb{R}^{dp}:|c_j|_2\leq R} \left\{ \sum_{t=1}^T \ell(\boldsymbol{c}, x_t) + 2|\boldsymbol{c}|_0 \sqrt{T} \log\left(1 + \frac{\sum_{j=1}^p |\boldsymbol{c}_j|_2}{\sqrt{6}|\boldsymbol{c}|_0 \tau}\right) \right\}$$
$$+ C\sqrt{T}\left(\sqrt{6pT}\tau + 2CR^2\right) + \frac{12pd\sqrt{T}\tau^2}{R^2}.$$

This result gives a sparsity regret bound when λ varies over time in the sequential procedure. If we choose a scale parameter $\tau = (MT)^{-1/2}$, it leads to a residual term of order $\mathcal{O}(\sqrt{T})$ as in Corollary 2. However, this choice is not possible in this adaptive setting of unknown horizon T. Note that this problem also occurs in [16], where a doubling trick is recommended to get a fully automatic algorithm.

4 Batch revisited

In this section, we go back to the standard clustering problem. Let us introduce an unknown probability P over the metric space $\mathbb{R}^d$, such that $\mathbb{E}_P|X|_2^2 < \infty$. Given an i.i.d. sample of random variable $X_1, \ldots, X_n$ drawn from P, and an integer $k \geq 1$, we want to find a set of k cluster's centers, or codebook $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$ which resume the law of P. Many authors have studied this problem. Usually, we want to minimize a distortion of the form:

$$\mathcal{W}_k(\mathbf{c}) = \mathbb{E}_P \min_{j=1,\ldots,k} |X - c_j|_2^2, \ \forall \mathbf{c} \in \mathbb{R}^{dk},$$

where $|\cdot|_2$ denotes the Euclidean distance in $\mathbb{R}^d$. In this setting, it is extremely standard to minimize an empirical risk based on the i.i.d. sample $X_1, \ldots, X_n$, defined as:

$$\widehat{\mathcal{W}}_k(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^n \min_{j=1,\ldots,k} |X_i - c_j|_2^2, \ \forall \mathbf{c} \in \mathbb{R}^{dk}. \tag{4.1}$$

The existence of a minimizer of (4.1) has been proved in [17]. The consistency as well as central limit theorem have been showed by Pollard (see [25] and [26]). However, in practice, we can note two principal drawbacks of this approach. Firstly, as mentionned in [7], it is not possible to reach the global minimum of (4.1), since we are faced to a non-convex minimization problem. Standard algorithms, such as the Lloyd algorithm, are made of Newton's type iterations and depend strongly on the initialization step. Moreover, the knowledge of k in the problem of clustering is not always guaranteed and a data-driven choice of this parameter remains a hard issue. In the following, we propose to use the Pac-Bayesian framework to get a fully automatic algorithm that performs model selection clustering.

Finally, for completeness, we also consider in this section the problem of high dimensional clustering. In this case, we suppose that the number of clusters k is known but the dimension d of the variable X could be much larger than the sample size n.

4.1 Model selection clustering

Recently, [15] formulates the problem of selecting the number of clusters k as a problem of model selection. She gives standard-style statistical learning bounds by using empirical process theory. For any integer $k \geq 1$, let us denote $\hat{\mathbf{c}}_k$ the minimizer of (4.1). Given the family $\{\hat{\mathbf{c}}_k, k = 1, \ldots, n\}$, [15] suggests a penalized model selection procedure to choose k as follows:

$$\hat{k} = \arg \min_{k=1,\ldots,n} \left\{ \widehat{\mathcal{W}}_k(\hat{\mathbf{c}}_k) + \mathrm{pen}_d(k) \right\},$$

where $\mathrm{pen}_d(k)$ is an increasing function of the dimension kd. In practice, the choice of the penalty is made in two steps:

1. A theoretical study gives the shape of the penalty, namely here (see [15][Theorem 2.1]):
$$\mathrm{pen}_d(k) = \square \sqrt{\frac{kd}{n}}, \text{ for some } \square > 0.$$
2. Then, the constant $\square > 0$ in front of the penalty's shape can be calibrated thanks to the slope heuristic (see [5]).

In this paragraph, we develop the Pac-Bayesian analysis of the previous sections as follows. Let us introduce an integer $p \geq 1$, which could be large enough (we can choose $p = n$ to fix the ideas). Consider the prior $\pi_S \in \mathcal{M}_1^+(\mathbb{R}^{dp})$ defined as (see (2.6)):

$$\pi_S(d\mathbf{c}) := \prod_{j=1}^{p} \left\{ C_{R,\tau}^{-1} \left(1 + \frac{|c_j|_2^2}{6\tau^2} \right)^{-\frac{3+d}{2}} \mathbb{I}(|c_j|_2 \leq 2R) \right\} d\mathbf{c}.$$

Fix $S_0 \equiv 0$ and draw $\hat{\mathbf{c}}_1$ according to π. Then, for any $i \in \{1, \ldots, n\}$:

- Get X_i and compute: $S_i(\mathbf{c}) = S_{i-1}(\mathbf{c}) + \ell(\mathbf{c}, X_i) + \frac{\lambda}{2}[\ell(\mathbf{c}, X_i) - \ell(\hat{\mathbf{c}}_i, X_i)]^2, \ \forall \mathbf{c} \in \mathbb{R}^{dp}$.
- Let $\hat{p}_{i+1}(dc) := \frac{e^{-\lambda S_i(\mathbf{c})}}{W_i} \pi(d\mathbf{c}) \in \mathcal{M}_1^+(\mathbb{R}^{dp})$ where $W_i = \mathbb{E}_{\mathbf{c}\sim\pi} e^{-\lambda S_i(\mathbf{c})}$.
- Draw $\hat{\mathbf{c}}_{i+1}$ according to $\hat{p}_{i+1}$.

The final estimator in the i.i.d. case, denoted as $\hat{\mathbf{c}}_{\mathrm{MA}}$, is a realization of the uniform law over $\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_{n+1}\}$:

$$\hat{\mathbf{c}}_{\mathrm{MA}} \sim \hat{\mu} := \mathcal{U}(\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_{n+1}\}), \tag{4.2}$$

where $\hat{\mu}$ is the uniform law over the set of estimators $\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_{n+1}\}$, conditionally to the training set $\mathcal{D}_n$. This additional step is called Miror Averaging (MA) and has been used in the i.i.d. setting by many authors (see for instance [20, 13, 2]). Since $\hat{\mathbf{c}}_{\mathrm{MA}}$ is a realization of an uniform law, we are finally interested in the expectation (with respect to the training set $\mathcal{D}_n$) of the expected risk of $\hat{\mathbf{c}}_{\mathrm{MA}}$, given by:

$$\mathbb{E}_{\mathcal{D}_n} \mathbb{E}_{\mathbf{c}'\sim\hat{\mu}} \mathcal{W}(\mathbf{c}') = \mathbb{E}_{\mathcal{D}_n} \frac{1}{n+1} \sum_{i=1}^{n+1} \mathcal{W}(\hat{\mathbf{c}}_i).$$

The main result of this paragraph is a penalized oracle inequality for the miror averaging estimator defined in (4.2).

Theorem 3 *Suppose the distribution P satisfies $P(|X|_2 \leq B) = 1$ for some $B > 0$. Let us consider the miror averaging $\hat{\mathbf{c}}_{\mathrm{MA}}$ defined in (4.2) using parameters $R > 0$, $p = n$ and prior π_S defined in (2.6). If we choose $(\tau, \lambda) = \left(n^{-1}, n^{-1/2}\right)$, the following holds:*

$$\mathbb{E}_{\mathcal{D}_n} E_{\boldsymbol{c}'\sim\hat{\mu}} \mathcal{W}(\boldsymbol{c}') \leq \inf_{1\leq k\leq n} \left\{ \mathcal{W}(\boldsymbol{c}_k^\star) + \frac{2k}{\sqrt{n}} \log\left(1 + \frac{n \sum_{j=1}^n |\boldsymbol{c}_j|_2}{\sqrt{6}k} \right) \right\} + n^{-1/2} \left(2R^2(2B + 3R)^2 + \sqrt{6}(2B + 3R) + \frac{12d}{R^2 n} \right),$$

where $\boldsymbol{c}_k^\star = \arg\min_{\boldsymbol{c}\in\mathbb{R}^{dn}:|\boldsymbol{c}|_0=k, |c_j|_2\leq R} \mathcal{W}(\boldsymbol{c})$.

The RHS of Theorem 3 can be compared with [15], where the penalized model selection procedure described above is used. The inequality of Theorem 3 ensures that in the i.i.d. case, the risk of our procedure is as well as the risk of the best codebook in the family, up to a residual term. This term approaches the rate $n^{-1/2}$, up to a $\log n$ factor.

From a model selection point of view, if we compare this result with [15], the main advantage of our approach is that there is not any tuning parameter to choose. The miror averaging estimator built in (4.2) reaches a penalized oracle inequality without any parameter to tune. This comes from the Pac-Bayesian analysis used in this paper.

4.2 High dimensional clustering

In this paragraph, we turn out into the problem of high dimensional clustering (see [6, 24]). Let us briefly introduce the model. Given an integer $k \geq 1$, we consider an i.i.d. sample $X_1, \ldots, X_n$ with unknown law P over $\mathbb{R}^d$, where d could be much larger than n. In this framework, we are interested in a codebook $\mathbf{c} = (c_1, \ldots, c_k)$ such that $|c_j|_0 << d$ for any $j = 1, \ldots, k$, where here, $|\cdot|_0$ stands for the usual ℓ_0-norm (i.e. the number of non-zero components in c_j). The main result of this paragraph is a sparsity oracle inequality for the miror averaging estimator defined in (4.2) with a slightlty different prior. In this setting of high dimensional clustering, we introduce the following sparsity prior:

$$\pi'_S(d\mathbf{c}) := \prod_{i=1}^{d} \left\{ C_{R,\tau}^{-1} \left(1 + \frac{|c_i|_2^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \mathbb{I}(|c_i|_2 \leq 2R) \right\} d\mathbf{c}, \tag{4.3}$$

where $c_i = (c_{i1}, \ldots, c_{ik}) \in \mathbb{R}^k$ denotes the vector of the i^{th} coordinates of each c_j in $\mathbf{c} = (c_1, \ldots, c_k)$, and $C_{R,\tau} := c_{k,\tau}\mathbb{P}(\sqrt{2}\tau T_k(3) \in \mathcal{B}_2(2R))$ for some constant $c_{k,\tau} > 0$. Let us briefly explain the introduction of this modified prior. Since we are looking at sparsity with respect to the dimension of the problem, we construct a product of d multivariate $T_k(3)$ Student's distribution, where $k \geq 1$ is the known number of clusters in the problem. This choice mimics the introduction of π_S in the model selection case. It encourages codebook $\mathbf{c}$ with small sparsity index $|\mathbf{c}|'_0$ defined as $|\mathbf{c}|'_0 = \text{card}\{i = 1, \ldots, d : \sum_{j=1}^{k} c_{ij}^2 \neq 0\}$.

Theorem 4 *Suppose distribution P satisfies $P(|X|_2 \leq B) = 1$. For some integer $k \geq 1$, let us consider the miror averaging $\hat{\mathbf{c}}_{\text{MA}}$ defined in (4.2) using prior π'_S defined in (4.3) with parameters $R > 0$. If we choose: $(\tau, \lambda) = \left((dn)^{-1/2}, n^{-1/2}\right)$, the following holds:*

$$\mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{\boldsymbol{c}'\sim\hat{\mu}} R(\boldsymbol{c}') \leq \inf_{\boldsymbol{c}\in\mathbb{R}^{dk}:\forall j, |c_j|_2 \leq R} \left\{ \mathcal{W}(\boldsymbol{c}) + \frac{2|\boldsymbol{c}|'_0}{\sqrt{n}} \log\left(1 + \frac{\sqrt{nd}\sum_{i=1}^{d}|\boldsymbol{c}_i|_2}{|\boldsymbol{c}|'_0}\right)\right\}$$
$$+ n^{-1/2}\left(\sqrt{6}(3R + 2B) + 2R^2(3R + 2B)^2 + \frac{12k}{R^2 n}\right).$$

where $|\boldsymbol{c}|'_0 = \text{card}\{i = 1, \ldots, d : \sum_{j=1}^{k} c_{ij}^2 \neq 0\}$ is the sparsity index of the codebook $\boldsymbol{c}$.

The RHS of Theorem 4 gives a rates of convergence of the form $\log d/\sqrt{n}$, which has to be compared with the usual rate $\log d/n$ in high dimensional linear regression. Here, the presence of a non-convex loss function gives rise to a rate of order $\mathcal{O}(n^{-1/2})$, up to a classical $\log d$ term.

5 Discussion

As a conclusion, we have taken a Pac-Bayesian point of view that I hope will shed some light on the issue of clustering. At the first glance, we consider the problem of online clustering of a deterministic sequence $x_t \in \mathbb{R}^d$, $t = 1, \ldots, T$. Interestingly, this problem could be seen as a prediction problem and allows to construct online clustering algorithms inspired from the prediction literature. Using a Pac-Bayesian analysis and a new kind of sparsity prior called

group-sparsity prior, we lead to sparsity regret bounds for our sequential algorithms. These results give the opportunity to deal with clustering of a non-stationnary process, i.e. with possible moving clusters and without any a priori on the number of clusters.

Finally, if we go back to the classical i.i.d. setting, clustering have been widely investigated. In this contribution, using pseudo-Bayesian estimators, we are abble to perform model selection clustering as well as high dimensional clustering. These two problems are related with a different kind of group-sparsity. It allows to perform them in a unified algorithm, by using two slightly different group-sparsity priors.

The main issue for future works is to give a practical way to compute the presented algorithms. In this direction, several authors suggest to use Markov chains with rare events to approximate appropriate Gibbs distributions on the prediction space. These recent techniques seem to be applicable in our setting. It could be a way of performing clustering in an online way, with many potential applications (see Section 1 for an illustration). Moreover, even in the classical batch setting, it could be an alternative to the classical k-means scenario, which has many drawbacks, such as the dependence on the initialization and the dependence on the number of clusters k.

6 Appendix

6.1 Proof of Lemma 1

First note that, for any $\mathbf{c} \in \mathbb{R}^{dp}$ such that $\forall j = 1, \dots, p,\ c_j \in \mathcal{B}_2(R)$, we have by definition of π_S and p_0:

$$\begin{aligned}\mathcal{K}(p_0, \pi_S) &= \int_{\mathbb{R}^{dp}} \log \left[\left(\frac{C_{R,\tau}}{C'_{R,\tau}} \right)^p \prod_{j=1}^p \frac{1 + |c'_j|_2^2/6\tau^2}{1 + |c'_j - c_j|_2^2/6\tau^2} \right] p_0(d\mathbf{c}') \\ &= p \log \frac{C_{R,\tau}}{C'_{R,\tau}} + \sum_{j=1}^p \int_{\mathbb{R}^{dp}} \log \left[\frac{1 + |c'_j|_2^2/6\tau^2}{1 + |c'_j - c_j|_2^2/6\tau^2} \right] p_0(d\mathbf{c}').\end{aligned}$$

To prove the first assertion, we hence have to show that:

$$\sum_{j=1}^p \int_{\mathbb{R}^{dp}} \log \left[\frac{1 + |c'_j|_2^2/6\tau^2}{1 + |c'_j - c_j|_2^2/6\tau^2} \right] p_0(d\mathbf{c}') \leq 2 \sum_{j=1}^p \log \left(1 + \frac{|c_j|_2}{\sqrt{6}\tau} \right). \tag{6.1}$$

By simple algebra, we have, for any $c, c' \in \mathbb{R}^d$, and any $a > 0$:

$$\frac{a^2 + |c'|_2^2}{a^2 + |c' - c|_2^2} = 1 + \frac{2a\langle c' - c, c/a \rangle}{a^2 + |c' - c|_2^2} \leq 1 + \frac{|c|_2}{a} + \frac{|c|_2^2}{a^2} \leq \left(1 + \frac{|c|_2}{a} \right)^2.$$

Applying this result for any $j = 1, \dots, p$ and for $a = \sqrt{6}\tau$ leads to:

$$\sum_{j=1}^p \int_{\mathbb{R}^{dp}} \log \left[\frac{1 + |c'_j|_2^2/6\tau^2}{1 + |c'_j - c_j|_2^2/6\tau^2} \right] p_0(d\mathbf{c}') \leq 2 \sum_{j=1}^p \log \left(1 + \frac{|c_j|_2}{\sqrt{6}\tau} \right).$$

For the second assertion, first note that, by concavity of $x \mapsto \log(1 + x)$, we have:

$$2 \sum_{j=1}^p \log \left(1 + \frac{|c_j|_2}{\sqrt{6}\tau} \right) \leq 2|\mathbf{c}|_0 \log \left(1 + \frac{\sum_{j=1}^p |c_j|_2}{\sqrt{6}\tau |\mathbf{c}|_0} \right).$$

Moreover, by definition of π_S and p_0, we have, for some random variable $\mathcal{T} \in \mathbb{R}^d$ with multivariate Student's distribution $T_d(3)$:

$$\begin{aligned}
\log\left(\frac{C_{R,\tau}}{C'_{R,\tau}}\right) &= \log\left(\frac{\mathbb{P}(\sqrt{2}\tau\mathcal{T} \in \mathcal{B}_2(2R))}{\mathbb{P}(\sqrt{2}\tau\mathcal{T} \in \mathcal{B}_2(R))}\right) \\
&\leq \log\left(1 + \frac{\mathbb{P}(|\mathcal{T}|_2 > (\sqrt{2}\tau)^{-1}R)}{\mathbb{P}(\sqrt{2}\tau\mathcal{T} \in \mathcal{B}_2(R))}\right) \\
&\leq \frac{\mathbb{P}(|\mathcal{T}|_2 > (\sqrt{2}\tau)^{-1}R)}{\mathbb{P}(\sqrt{2}\tau\mathcal{T} \in \mathcal{B}_2(R))} \\
&\leq \frac{2\tau^2\mathbb{E}|\mathcal{T}|_2^2}{R^2\mathbb{P}(\sqrt{2}\tau\mathcal{T} \in \mathcal{B}_2(R))} \\
&\leq \frac{4\tau^2\mathbb{E}|\mathcal{T}|_2^2}{R^2} = \frac{12d\tau^2}{R^2},
\end{aligned}$$

provided that $R \geq \sqrt{2}\tau t_d(1/2)$, where $t_d(1/2)$ is such that $\mathbb{P}(\mathcal{T} \in \mathcal{B}_2(t_d(1/2)) = 1/2$. Then, since $|\mathcal{T}|_2^2/d$ has a Fisher distribution $\mathcal{F}(d,3)$ (see [21]), the last equality follows easily.

□

6.2 Proof of Corollary 1 and Corollary 2

The proof of Corollary 1 is a direct consequence of Theorem 1 and the introduction of the sparsity prior of Lemma 1. Let us consider $\tau, R > 0$, $\mathbf{c} \in \mathbb{R}^{dp}$ such that $c_j \in \mathcal{B}_2(R)$ for any $j \in \{1,\dots,p\}$. Denote p_0 the translated version of π_S with mean $\mathbf{c}$. By using Lemma 1, gathering with Theorem 1, we have:

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}_{(\hat{p}_1,\dots,\hat{p}_t)}\ell(\hat{\mathbf{c}}_t, x_t) \leq \sum_{t=1}^T \ell(\mathbf{c}, x_t) + \mathbb{E}_{\mathbf{c}'\sim p_0}\sum_{t=1}^T[\ell(\mathbf{c}', x_t) - \ell(\mathbf{c}, x_t)] \\
+ \frac{2|\mathbf{c}|_0}{\lambda}\log\left(1 + \frac{\sum_{j=1}^p |c_j|_2}{\sqrt{6}|\mathbf{c}|_0\tau}\right) + \frac{12pd\tau^2}{\lambda R^2} + \frac{\lambda}{2}\mathbb{E}_{(\hat{p}_1,\dots,\hat{p}_T)}\mathbb{E}_{\mathbf{c}'\sim p_0}\sum_{t=1}^T[\ell(\mathbf{c}', x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2.
\end{aligned}$$

Now, for any $t \in \{1,\dots,T\}$, by definition of $\mathbf{c}$ and p_0, we have p_0-a.s. :

$$\begin{aligned}
|\ell(\mathbf{c}, x_t) - \ell(\mathbf{c}', x_t)| &\leq \max_{j=1,\dots,p}\left||x_t - c_j|_2^2 - |x_t - c'_j|_2^2\right| \\
&\leq (3R + 2B_T)\max_{j=1,\dots,p}|c_j - c'_j|_2,
\end{aligned}$$

where with a slight abuse of notations, $|\cdot|_2$ stands for the Euclidean norm in $\mathbb{R}^{dp}$. Then, gathering with the previous inequality, we arrive at:

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}_{(\hat{p}_1,\dots,\hat{p}_t)}\ell(\hat{\mathbf{c}}_t, x_t) \leq \sum_{t=1}^T \ell(\mathbf{c}, x_t) + \frac{2|\mathbf{c}|_0}{\lambda}\log\left(1 + \frac{\sum_{j=1}^p |c_j|_2}{\sqrt{6}|\mathbf{c}|_0\tau}\right) + CT\mathbb{E}_{\mathbf{c}'\sim p_0}\max_{j=1,\dots,p}|c'_j - c_j|_2 \\
+ \frac{\lambda}{2}C^2T\mathbb{E}_{(\hat{p}_1,\dots,\hat{p}_T)}\mathbb{E}_{\mathbf{c}'\sim p_0}\max_{j=1,\dots,p}|\hat{c}_j - c'_j|_2^2 + \frac{12pd\tau^2}{\lambda R^2},
\end{aligned}$$

where $C := 3R + 2B_T$. Last step is to control the last two terms in the previous inequality. By definition of the measure p_0 and a standard maximal inequality (see for instance [28]), we have:

$$\mathbb{E}_{\mathbf{c}'\sim p_0}\max_{j=1,\dots,p}|c'_j - c_j| \leq \sqrt{p}\max_{j=1,\dots,p}\mathbb{E}_{\mathbf{c}'\sim p_0}|c'_j - c_j| \leq \sqrt{6p}\tau.$$

The last term can be controlled as follows:

$$\mathbb{E}_{(\hat{p}_1,\dots,\hat{p}_T)}\mathbb{E}_{\mathbf{c}'\sim p_0}\max_{j=1,\dots,p}|\hat{c}_j - c'_j|^2 \leq 4R^2.$$

Then, Corollary 1 is proved. Corollary 2 follows easily with a proper choice of $\tau, \lambda > 0$.

6.3 Proof of Theorem 2

The proof is based on an adaptive version of the Pac-Bayesian bound of Theorem 1. Indeed, we can show that the adaptive algorithm with sequence $(\lambda_t)_{t=1}^{T+1}$ satisfies the following bound:

Proposition 1 *For any deterministic sequence $(x_t)_{t=1}^T$, for any prior π, for any non-increasing sequence $(\lambda_t)_{t=1,\dots,T+1}$, the cumulative loss of the adaptive algorithm satisfies:*

$$\mathcal{E}_T(\hat{\boldsymbol{c}}) \leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\boldsymbol{c}\sim\rho} \sum_{t=1}^T \ell(\boldsymbol{c}, x_t) + \frac{\mathcal{K}(\rho,\pi)}{\lambda_{T+1}} + \frac{\lambda_T}{2} \mathbb{E}_{(\hat{p}_1,\dots,\hat{p}_T)} \mathbb{E}_{\boldsymbol{c}\sim\rho} \sum_{t=1}^T [\ell(\boldsymbol{c}, x_t) - \ell(\hat{\boldsymbol{c}}_t, x_t)]^2 \right\}.$$

The idea of the proof is to control the quantity $\log(W_T)/\lambda_{T+1} - \log(W_0)/\lambda_1$. On the one side, using the definition of W_T, we have:

$$\begin{aligned} \frac{\log W_T}{\lambda_{T+1}} - \frac{\log W_0}{\lambda_1} &= \frac{1}{\lambda_{T+1}} \log \left(\mathbb{E}_{\mathbf{c}\sim\pi} e^{-\lambda_{T+1} S_T(\mathbf{c})} \right) \\ &= - \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\mathbf{c}\sim\rho} S_T(\mathbf{c}) + \frac{\mathcal{K}(\rho,\pi)}{\lambda_{T+1}} \right\}, \end{aligned} \tag{6.2}$$

where the second equality comes from the duality formula (2.3). On the other side, by cancellation of sum argument, we can write:

$$\begin{aligned} \frac{\log W_T}{\lambda_{T+1}} - \frac{\log W_0}{\lambda_1} &= \sum_{t=1}^T \left(\frac{\log W_t}{\lambda_{t+1}} - \frac{\log W_{t-1}}{\lambda_t} \right) \\ &= \sum_{t=1}^T \left(\frac{\log W_t}{\lambda_{t+1}} - \frac{\log W_t'}{\lambda_t} + \frac{1}{\lambda_t} \log \frac{W_t'}{W_{t-1}} \right), \end{aligned} \tag{6.3}$$

where W_t' is defined for any $t = 1, \dots, T$ as:

$$W_t' = \mathbb{E}_{\mathbf{c}\sim\pi} e^{\lambda_t S_t(\mathbf{c})}.$$

Now, note that by Jensen's inequality, we have:

$$\frac{\log W_t}{\lambda_{t+1}} = \frac{\log \left(\mathbb{E}_{\mathbf{c}\sim\pi} \left[\left(e^{-\lambda_t S_t(\mathbf{c})} \right)^{\lambda_{t+1}/\lambda_t} \right] \right)}{\lambda_{t+1}} \leq \frac{\log \left(\left[\mathbb{E}_{\mathbf{c}\sim\pi} e^{-\lambda_t S_t(\mathbf{c})} \right]^{\lambda_{t+1}/\lambda_t} \right)}{\lambda_{t+1}} = \frac{\log W_t'}{\lambda_t}.$$

Then, one obtains:

$$\frac{\log W_t}{\lambda_{t+1}} - \frac{\log W_t'}{\lambda_t} \leq 0. \tag{6.4}$$

We turn out into the last term in (6.3). We have, by definition of $\hat{p}_t$:

$$\begin{aligned} \frac{1}{\lambda_t} \log \frac{W_t'}{W_{t-1}} &= \frac{1}{\lambda_t} \log \left(\frac{\mathbb{E}_{\mathbf{c}\sim\pi} e^{-\lambda_t S_t(\mathbf{c})}}{\mathbb{E}_{\mathbf{c}\sim\pi} e^{-\lambda_t S_{t-1}(\mathbf{c})}} \right) \\ &= \frac{1}{\lambda_t} \log \left(\frac{\int_{\mathbb{R}^{dp}} e^{-\lambda_t [\ell(\mathbf{c},x_t) + \frac{\lambda_t}{2}[\ell(\mathbf{c},x_t) - \ell(\hat{\mathbf{c}}_t,x_t)]^2]} e^{-\lambda_t S_{t-1}(\mathbf{c})} d\pi(\mathbf{c})}{\int_{\mathbb{R}^{dp}} e^{-\lambda_t S_{t-1}(\mathbf{c})} d\pi(\mathbf{c})} \right) \\ &= \frac{1}{\lambda_t} \log \left(\mathbb{E}_{\mathbf{c}\sim\hat{p}_t} e^{-\lambda_t [\ell(\mathbf{c},x_t) + \frac{\lambda_t}{2}[\ell(\mathbf{c},x_t) - \ell(\hat{\mathbf{c}}_t,x_t)]^2]} \right). \end{aligned} \tag{6.5}$$

Now, using (2.4), we can write:

$$\mathbb{E}_{\mathbf{c}\sim\hat{p}_t} e^{-\lambda_t [\ell(\mathbf{c},x_t) + \frac{\lambda_t}{2}[\ell(\mathbf{c},x_t) - \ell(\hat{\mathbf{c}}_t,x_t)]^2]} \leq -\mathbb{E}_{\mathbf{c}\sim\hat{p}_t} \ell(\mathbf{c}, x_t).$$

Applying the above inequality to (6.5), summing over t and integrating with respect to $(\hat{p}_1, \ldots, \hat{p}_T)$, we arrive at:

$$\mathbb{E}_{\hat{p}_1^T}\left[\frac{\log W_T}{\lambda_{T+1}} - \frac{\log W_0}{\lambda_1}\right] \leq -\sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1,\ldots,\hat{p}_t)} \ell(\hat{\mathbf{c}}_t, x_t). \tag{6.6}$$

Gathering with (6.2), we arrive at the conclusion of Proposition 1. The proof of Theorem 2 follows easily using the same paths as in the proof of Corollary 2. Let us consider $\tau, R > 0$, p_0 the measure defined in Lemma 1. As in the proof of Corollary 1, we can show with Proposition 1 that the adaptive algorithm satisfies:

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1,\ldots,\hat{p}_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \frac{2|\mathbf{c}|_0}{\lambda_{T+1}} \log\left(1 + \frac{\sum_{j=1}^{p} |c_j|_2}{\sqrt{6}|\mathbf{c}|_0 \tau}\right) + CT\mathbb{E}_{\mathbf{c}'\sim p_0} \max_{j=1,\ldots,p} |c'_j - c_j|_2 \\ + \frac{\lambda_T}{2} C^2 T \mathbb{E}_{(\hat{p}_1,\ldots,\hat{p}_T)} \mathbb{E}_{\mathbf{c}'\sim p_0} \max_{j=1,\ldots,p} |\hat{c}_j - c'_j|_2^2 + \frac{12pd\tau^2}{\lambda_{T+1} R^2},\end{aligned}$$

where $C = 3R + 2B_T$. The control of the RHS in the previous inequality follows exactly the proof of Corollary 1. It leads to the result when $\lambda_t = 1/\sqrt{t}$.

6.4 Proof of Theorem 3

The proof of Theorem 3 is based on an i.i.d. version of Theorem 1. Indeed, by using Theorem 3.1 in [2], it is clear that the miror averaging (4.2) satisfies the following PAC-Bayesian bound:

$$\mathbb{E}_{\mathcal{D}_n} \mathbb{E}_{\mathbf{c}'\sim\hat{\mu}} R(\mathbf{c}') \leq \min_{\rho\in\mathcal{M}_1^+(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\mathbf{c}\sim\rho} R(\mathbf{c}) + \frac{\mathcal{K}(\rho,\pi)}{\lambda(n+1)} + \frac{\lambda}{2} \mathbb{E}_{\mathcal{D}_n} \mathbb{E}_{\mathbf{c}\sim\rho} \mathbb{E}_{\mathbf{c}'\sim\hat{\mu}} \mathbb{E}_P [\ell(\mathbf{c}, X) - \ell(\mathbf{c}', X)]^2 \right\}, \tag{6.7}$$

where $\hat{\mu} := \mathcal{U}(\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_{n+1}\})$ is defined in (4.2). Now, using the same path as in the proof of Corollary 1, and using the fact that $|X|_\infty \leq B$, we have:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n} \mathbb{E}_{\mathbf{c}'\sim\hat{\mu}} R(\mathbf{c}') \leq \inf_{\mathbf{c}\in\mathbb{R}^{dp}:\forall j\ |c_j|\leq R} \left\{ R(\mathbf{c}) + \frac{2|\mathbf{c}|_0}{\lambda(n+1)} \log\left(1 + \frac{\sum_{j=1}^{p} |c_j|_2}{\sqrt{6}|\mathbf{c}|_0\tau}\right)\right\} \\ + \frac{12pd\tau^2}{\lambda(n+1)R^2} + (3R+2B)\sqrt{6p}\tau + 2R^2(3R+2B)^2\lambda.\end{aligned}$$

For $p = n$, the choice of (τ, λ) in Theorem 3 ends up the proof.

6.5 Proof of Theorem 4

The proof of Theorem 4 follows the proof of Theorem 3. The use of the modified sparsity prior π'_S defined in 4.3 gives us:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n} \mathbb{E}_{\mathbf{c}'\sim\hat{\mu}} R(\mathbf{c}') \leq \inf_{\mathbf{c}\in\mathbb{R}^{dk}:|c_j|\leq R} \left\{ R(\mathbf{c}) + \frac{2|\mathbf{c}|'_0}{\lambda(n+1)} \log\left(1 + \frac{\sum_{j=1}^{d} |c_j|_2}{\sqrt{6}|\mathbf{c}|'_0\tau}\right)\right\} \\ + \frac{12dk\tau^2}{\lambda(n+1)R^2} + (3R+2B)\sqrt{6d}\tau + 2R^2(3R+2B)^2\lambda.\end{aligned}$$

where $|\mathbf{c}|'_0 = \text{card}\{i = 1, \ldots, d : \sum_{j=1}^{k} c_{ij}^2 \neq 0\}$ is the sparsity index of the codebook $\mathbf{c}$. The choice of (τ, λ) is Theorem 4 gives:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n} \mathbb{E}_{\mathbf{c}'\sim\hat{\mu}} R(\mathbf{c}') \leq \inf_{\mathbf{c}\in\mathbb{R}^{dk}:|c_j|\leq R} \left\{ R(\mathbf{c}) + \frac{2|\mathbf{c}|'_0}{\sqrt{n}} \log\left(1 + \frac{\sqrt{nd}\sum_{i=1}^{d} |c_i|_2}{|\mathbf{c}|'_0}\right)\right\} \\ + \frac{12k}{R^2 n^{3/2}} + n^{-1/2}\left(\sqrt{6}(3R+2B) + 2R^2(3R+2B)^2\right).\end{aligned}$$

References

[1] P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14:243–280, 2013.

[2] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37 (4):1591–1646, 2009.

[3] P. Auer, N. Cesa-Bianci, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:48–75, 2002.

[4] A. Barron. Are bayes rules consistent in information ? In *Open Problems in Communication and Computation*, pages 85–91. Springer New-York, 1987.

[5] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22:455–470, 2012.

[6] C. Bouveyron and C. Brunet. Model-based clustering of high-dimensional data: A review. Computational Statistics and Data Analysis, to appear, 2013.

[7] S. Bubeck. How the initialization affects the k means. *IEEE Tansactions on Information Theory*, 48:2789–2793, 2002.

[8] O. Catoni. *Statistical Learning Theory and Stochastic Optimization.* Springer, Lecture Notes in Mathematics, 2001.

[9] N. Cesa-Bianchi and G. Lugosi. *Learning, Prediction and Games.* Cambridge University Press, 2006.

[10] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

[11] A. Choromanska and C. Monteleoni. Online clustering with experts. In Journal of Machine Learning Research (JMLR) Workshop and Conference Proceedings, editors, *Proceedings of ICML 2011 Workshop on Online Trading of Exploration and Exploitation 2*, 2012.

[12] A. S. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72 (1-2):39–61, 2008.

[13] A. S. Dalalyan and A.B. Tsybakov. Miror averaging with sparsity priors. *Bernoulli*, 18 (3):914–944, 2012.

[14] A. S. Dalalyan and A.B. Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *J. Comput. System Sci.*, 78:1423–1443, 2012.

[15] A. Fischer. On the number of groups in clustering. *Statistics and Probability Letters*, 81:1771–1781, 2011.

[16] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14:729–769, 2013.

[17] Siegfried Graf and Harald Luschgy. *Foundation of quantization for probability distributions.* Springer-Verlag, 2000. Lecture Notes in Mathematics, volume 1730.

[18] B. Guedj and P. Alquier. Pac-bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013.

[19] D. Haussler, J. Kivinen, and M. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44 (5):1906–1925, 1998.

[20] A. Juditsky, P. Rigollet, and A.B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36 (5):2183–2206, 2008.

[21] S. Kotz and S. Nadarajah. *Multivariate t distribution and their applications*. Cambridge University Press, 2004.

[22] N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[23] D.A. Mac Allester. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234. ACM, 1998.

[24] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: A review. *SIGKDD Explor. Newsl.*, 6(1):90–105, June 2004.

[25] D. Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9 (1), 1981.

[26] D. Pollard. A central limit theorem for k-means clustering. *The Annals of Probability*, 10 (4), 1982.

[27] M.W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

[28] A. W. van der Vaart and J. A. Wellner. *Weak convergence and Empirical Processes. With Applications to Statistics*. Springer Verlag, 1996.

[29] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68 (1):49–67, 2007.

[30] Shaoting Zhang, Junzhou Huang, Yuchi Huang, Yang Yu, Hongsheng Li, and Dimitris N. Metaxas. Automatic image annotation using group sparsity. In *In CVPR*, 2010.

[31] S. Zong. Efficient online spherical k-means clustering. In IEEE, editor, *Proceedings of International Joint Conference on Neural Networks IJCNN'05*, pages 3180–3185, 2005.